CLAIMS

WE CLAIM:

1. A method for classifying electronically posted documents, the method comprising:

receiving a first document and a second document;

generating a first metadata summary corresponding to said first document and a second metadata summary corresponding to the second document, wherein the first metadata summary includes a first summary sub-tree and the second metadata summary includes a second summary sub-tree;

comparing the structure of the first summary sub-tree with the structure of the second summary sub-tree; and

identifying the first and second documents as distinct if the structures of the first and second summary sub-trees are not equivalent.

2. The method of claim 1, wherein the first summary sub-tree includes at least one attribute having a first attribute value, and wherein the second summary sub-tree includes at least one attribute having a second attribute value, the method further comprising:

comparing, for each of the at least one attributes, the first and second attribute values; and

identifying the first and second documents as distinct if the attribute values of the first and second summary sub-trees are not equivalent.

3. The method of claim 1, wherein the first summary sub-tree includes text content, and wherein the second summary sub-tree includes text content, the method further comprising:

comparing the text content included within the first and second summary sub-trees; and

identifying the first and second documents as distinct if the text content of the first and second summary sub-trees are not equivalent.

4. The method of claim 2, wherein the first summary sub-tree further includes text content, and wherein the second summary sub-tree includes text content, the method further comprising:

comparing the text content included within the first and second summary sub-trees; and

identifying the first and second documents as distinct if the text content included within the first and second summary sub-trees are not equivalent.

5. The method of claim 4, further comprising identifying the first and second documents as duplicates if the text content within the first and second summary sub-trees are equivalent.

6. The method of claim 5, further comprising removing the second metadata summary from the first summary group if the structures of the first and second summary sub-trees are equivalent and if the first summary value is equivalent to the second summary value for each of the at least one attributes.

7. The method of claim 1, further comprising:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the structures of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

8. The method of claim 2, further comprising:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the attribute values of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

9. The method of claim 3, further comprising:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the text content of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

10. A method for classifying electronically posted documents, the method comprising:

receiving a plurality of documents;

generating a respective plurality of metadata summaries corresponding to the plurality of received documents;

grouping a first subset of the respective plurality of metadata summaries into a first summary group, the first summary group comprising summaries having a first mime-type designation;

selecting a first metadata summary and a second metadata summary from the first summary group, wherein the first metadata summary includes a first summary sub-tree and the second metadata summary includes a second summary sub-tree;

comparing the structure of the first summary sub-tree with the structure of the second summary sub-tree; and

identifying the first and second documents as distinct if the structures of the first and second summary sub-trees are not equivalent.

11. The method of claim 10, wherein grouping further comprises grouping a second subset of the respective metadata summaries into a second summary group, the second summary group comprising summaries having a second mime-type designation.

12. A system for classifying electronically posted documents, the system comprising:

a metadata parser module coupled to receive electronically posted documents, the metadata parser configured to output respective metadata summaries, wherein each respective metadata summary comprises one or more sub-trees structures, one or more attributes, and content text;

a summary repository coupled to receive and store the respective metadata summaries; and

a summary consolidator coupled to the summary repository, the summary consolidator configured to delete duplicate metadata summaries from the summary repository.

13. The system of claim 12, wherein the summary consolidator comprises:

a sub-tree comparator configured to compare one or more sub-tree structures of the retrieved metadata summaries;

an attribute comparator configured to compare the attribute values of the retrieved metadata summaries; and

a text comparator configured to compare the text content included within the retrieved metadata summaries.

14. The system of claim 13, wherein the sub-tree comparator is configured to compare the metadata portion of the metadata summary.

15. The system of claim 13, wherein the attribute comparator is configured to compare the attribute values included within the metadata portion of the metadata summary.

16. The system of claim 13, wherein the text comparator is configured to compare the text content included within the metadata portion of the metadata summary.

17. A program product for use in a computer system that executes program steps recorded in a computer-readable media to perform a method for classifying electronically posted documents, the program product comprising:

a recordable media;

a program of computer-readable instructions executable by the computer system to perform processes comprising:

receiving a first document and a second document;

generating a first metadata summary corresponding to said first document and a second metadata summary corresponding to the second document, wherein the first metadata summary includes a first summary sub-tree and the second metadata summary includes a second summary sub-tree;

comparing the structure of the first summary sub-tree with the structure of the second summary sub-tree; and

identifying the first and second documents as distinct if the structures of the first and second summary sub-trees are not equivalent.

18. The program product of claim 17, wherein the first summary sub-tree includes at least one attribute having a first attribute value, and wherein the second summary sub-tree includes at least one attribute having a second attribute value, the program product method further comprising the processes of:

comparing, for each of the at least one attributes, the first and second attribute values; and

identifying the first and second documents as distinct if the attribute values of the first and second summary sub-trees are not equivalent.

19. The program product of claim 18, wherein the first summary sub-tree includes text content, and wherein the second summary sub-tree includes text content, the program product further comprising the processes of:

comparing the text content included within the first and second summary sub-trees; and

identifying the first and second documents as distinct if the text content of the first and second summary sub-trees are not equivalent.

20. The program product of claim 19, further comprising the method step of identifying the first and second documents as duplicates if the text content within the first and second summary sub-trees are equivalent.

21. The program product of claim 20, further comprising the process of removing the second metadata summary from the first summary group.

22. The program product of claim 21, further comprising the processes of:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the text content of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

23. The method of claim 18, further comprising the processes of:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the attribute values of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

24. The method of claim 19, further comprising the processes of:

defining a first equivalence metadata table comprising:

a first row corresponding to the first metadata summary;

a second row corresponding to the second metadata summary;

a first column corresponding to the first metadata summary; and

a second column corresponding to the second metadata summary, wherein the process of identifying the first and second documents as distinct if the text content of the first and second summary sub-trees are not equivalent comprises storing a zero binary value in the first row and second column position of the equivalence metadata summary.

25. A program product for use in a computer system that executes program steps recorded in a computer-readable media to perform a method for classifying electronically posted documents, the program product comprising:

a recordable media;

a program of computer-readable instructions executable by the computer system to perform method steps comprising:

receiving a plurality of documents;

generating a respective plurality of metadata summaries corresponding to the plurality of received documents;

grouping a first subset of the respective plurality of metadata summaries into a first summary group, the first summary group comprising summaries having a first mime-type designation;

selecting a first metadata summary and a second metadata summary from the first summary group, wherein the first metadata summary includes a first summary sub-tree and the second metadata summary includes a second summary sub-tree;

comparing the structure of the first summary sub-tree with the structure of the second summary sub-tree; and

identifying the first and second documents as distinct if the structures of the first and second summary sub-trees are not equivalent.

26. The program product of claim 25, wherein the step of grouping further comprises the step of grouping a second subset of the respective metadata summaries into a second summary group, the second summary group comprising summaries having a second mime-type designation.